# Chordal Graphs in Computational Biology – New Insights and Applications

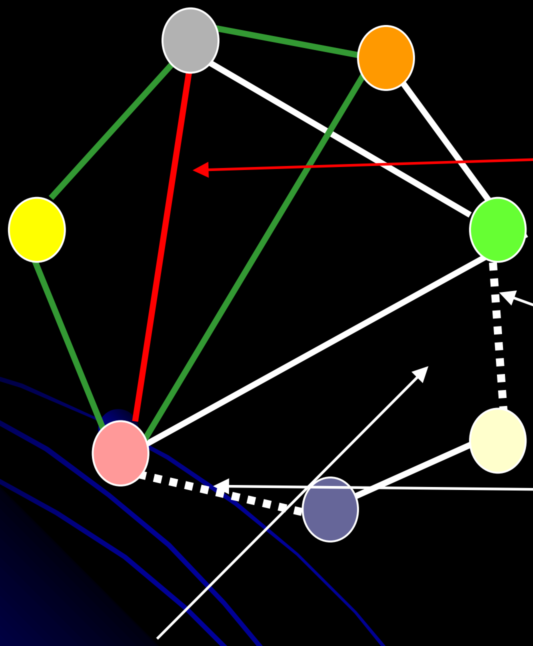
Teresa Przytycka

NIH / NLM / NCBI

# Overview

- Chordal graphs - *definitions and properties*
- Classical application to  perfect phylogeny
- New applications
  - Intron evolution
  - Understanding evolution of multi-domain proteins
  - Static and dynamic decomposition of protein complexes
- Conclusions

# Chordal graphs



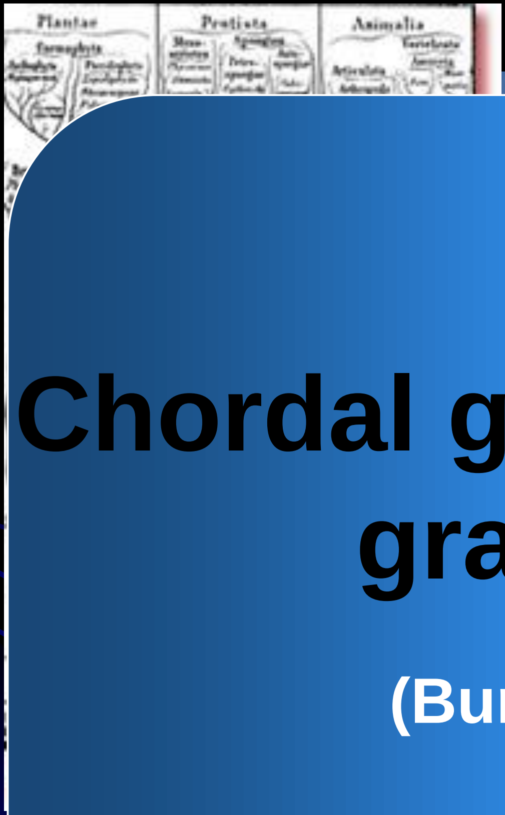
Chord = an edge connecting two non-consecutive nodes of a cycle

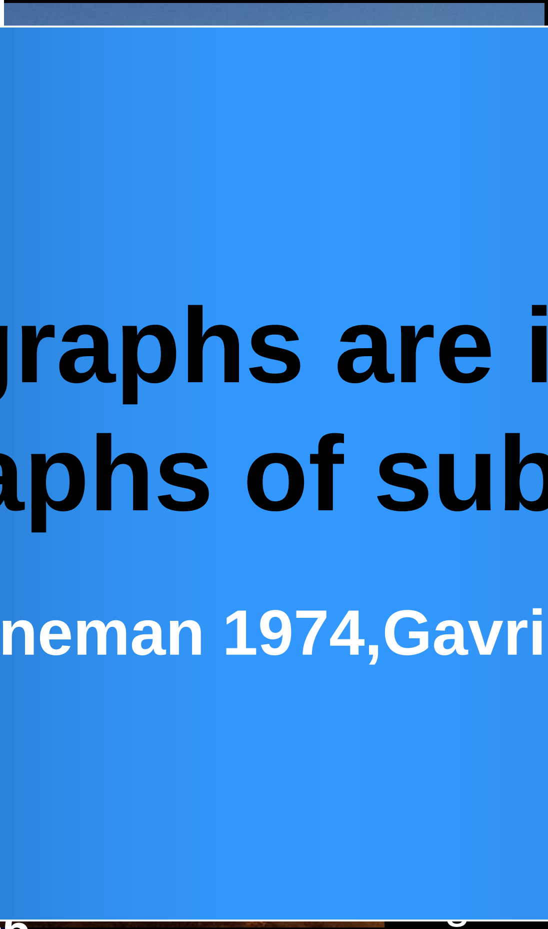
**Chordal graph** – every cycle of length at least four has a chord.

With these two edges the graph is not chordal

hole

# Applications to biology are prompted by the relation of chordal graphs to trees





**Chordal graphs are intersection graphs of subtrees**

**(Buneman 1974,Gavril 1974 )**

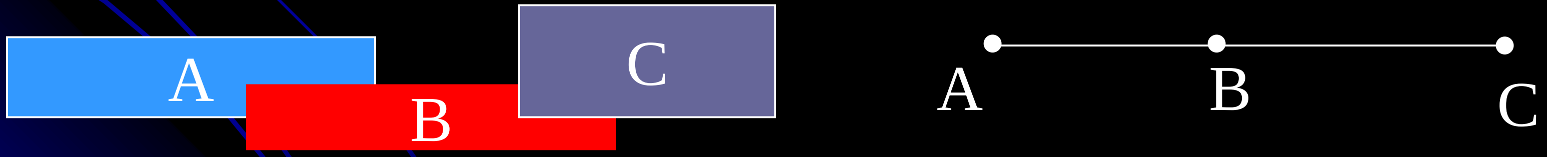
**T**
**the**
**Ernst Haeckel in 1866**

g

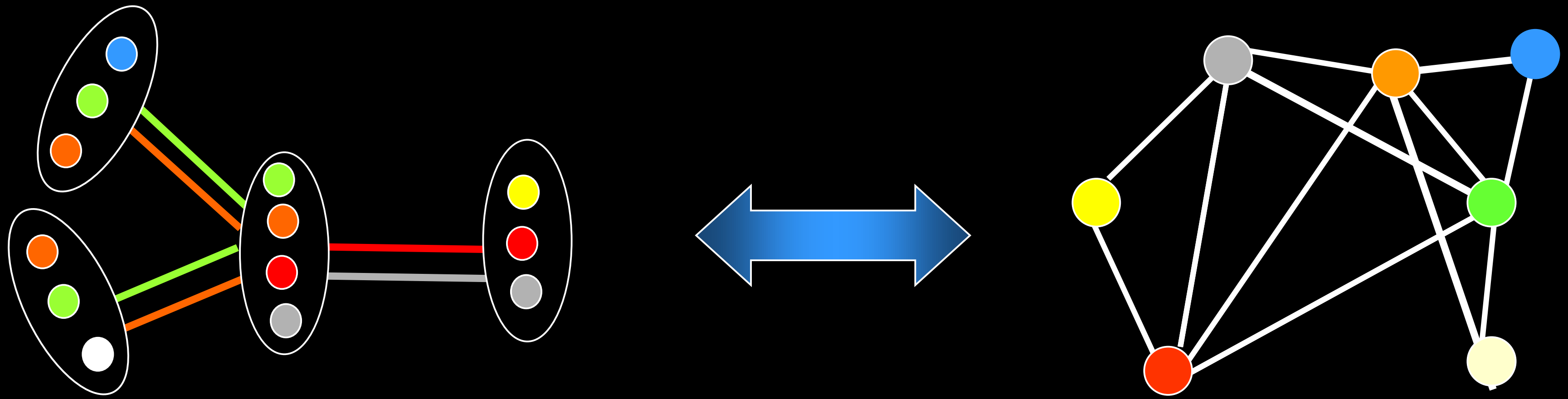
aul Berg

# Intersection graphs

- Nodes correspond to some objects (e.g. geometrical objects like rectangles on a plane)
- There is an edge between two such nodes if the corresponding objects intersect (share points)



# Intersection graphs of subtrees of a tree



intersection tree representation

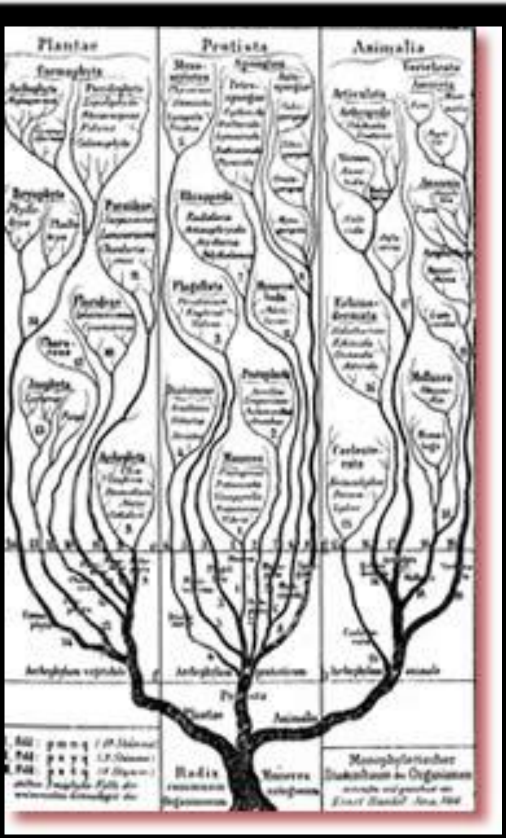
**Clique tree**:

Nodes = maximal cliques

For every graph node – the cliques containing this node span a sub-tree in the clique tree

Polynomial time algorithms **(Tarjan, Yannakakis, 1984)**

# Classical application of chordal graphs to evolutionary biology



# Taxa and characters

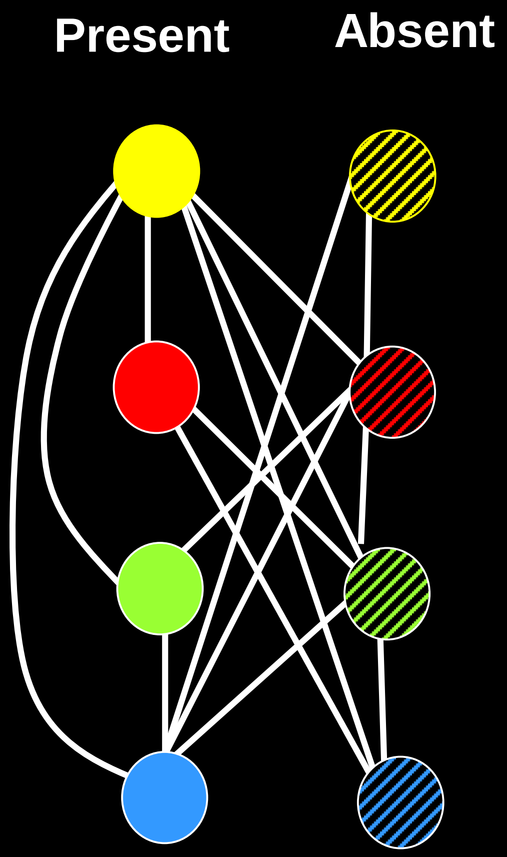
- Taxa set of biological entities that are evolutionarily related
- Each taxon is described by a set of characters which are subject to evolutionary changes
- Changes
  - Binary - two states 1/0 changes: insertions and deletions

# Constructing phylogenic tree

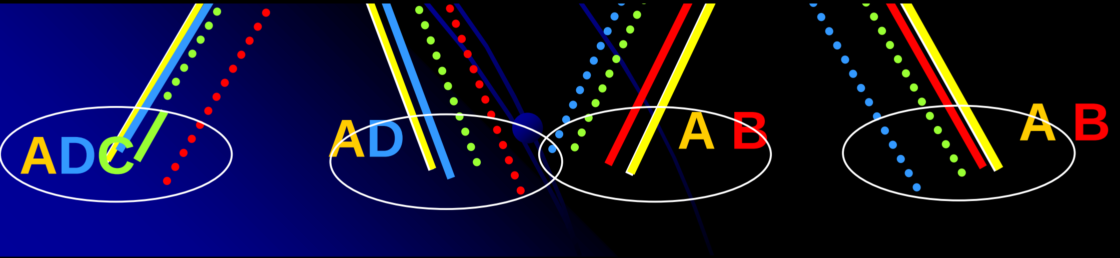
- **Using compatibility criterion**
- **Using maximum parsimony criterion**

# Perfect Phylogeny

Given the attributes of observed taxa is it possible to explain them by a perfect phylogeny tree?

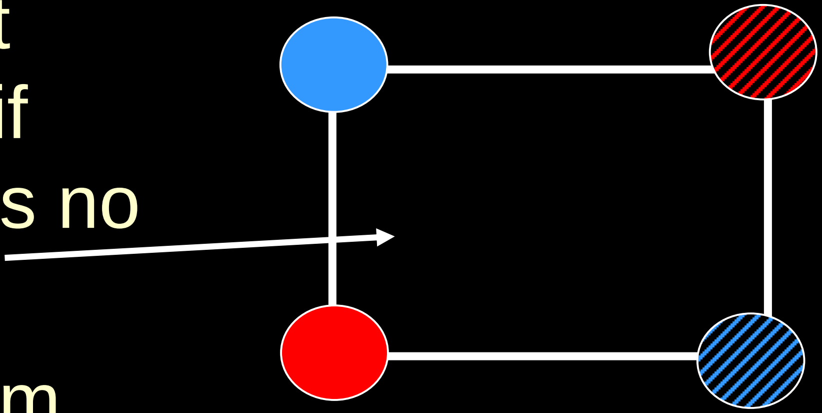


Attribute overlap graph



# Character Compatibility for binary characters

- A set of taxa admits perfect phylogeny if and only only if attributes overlap graph has no hole of this type
- Two characters are that form such hole are called non-



**Constructing phylogenic tree using compatibility criterion:**

**•Remove smallest number of characters so that the remaining characters are compatible**

**•Use the remaining characters to compute the tree**

(NP-complete)

# Parsimony methods for inferring phylogeny

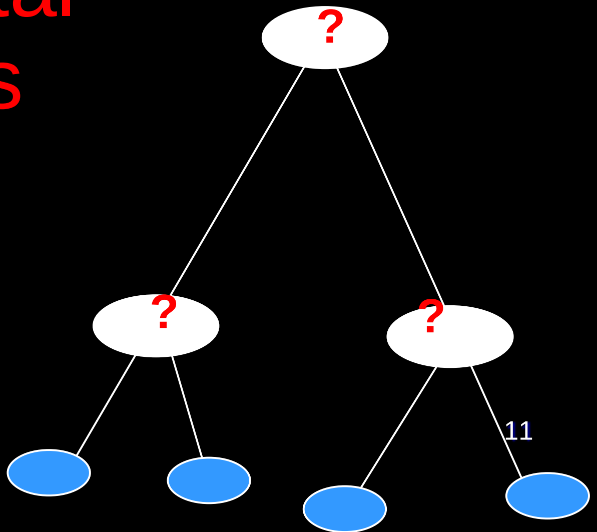
Build a tree such that

the input taxa  is in  the leaves

the inferred ancestral taxa in the internal nodes

and the attributes of the ancestral taxa are selected such that the total number of character changes

along edges is minimized.



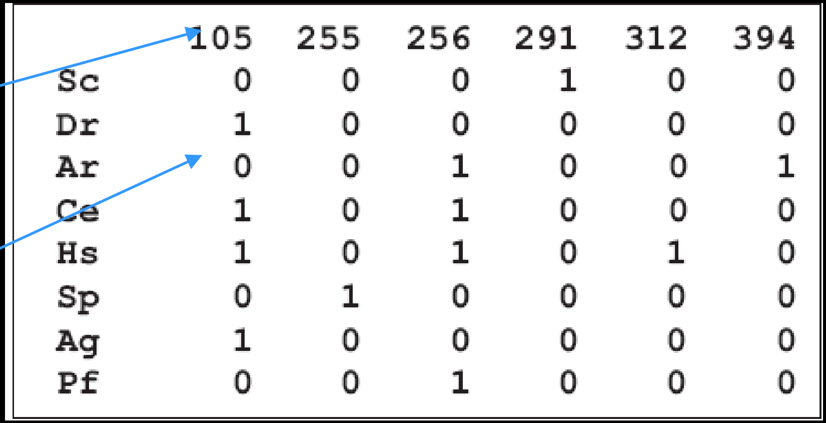
# Dollo parsimony

- Only one insertion per character
- Multiple deletions possible
- Appropriate for complex characters that are hard to gain but possible to lose

# Introns: Non coding sequences interrupting coding sequence in a gene

Introns:

- Independent insertion at the same position is unlikely
- Deletion possible
- Dollo parsimony seems reasonable
- Data assembled by Rogozin et al 2003
  - Multiple sequence alignment  orthologous genes
  - Identify intron start positions
  - Build binary table:

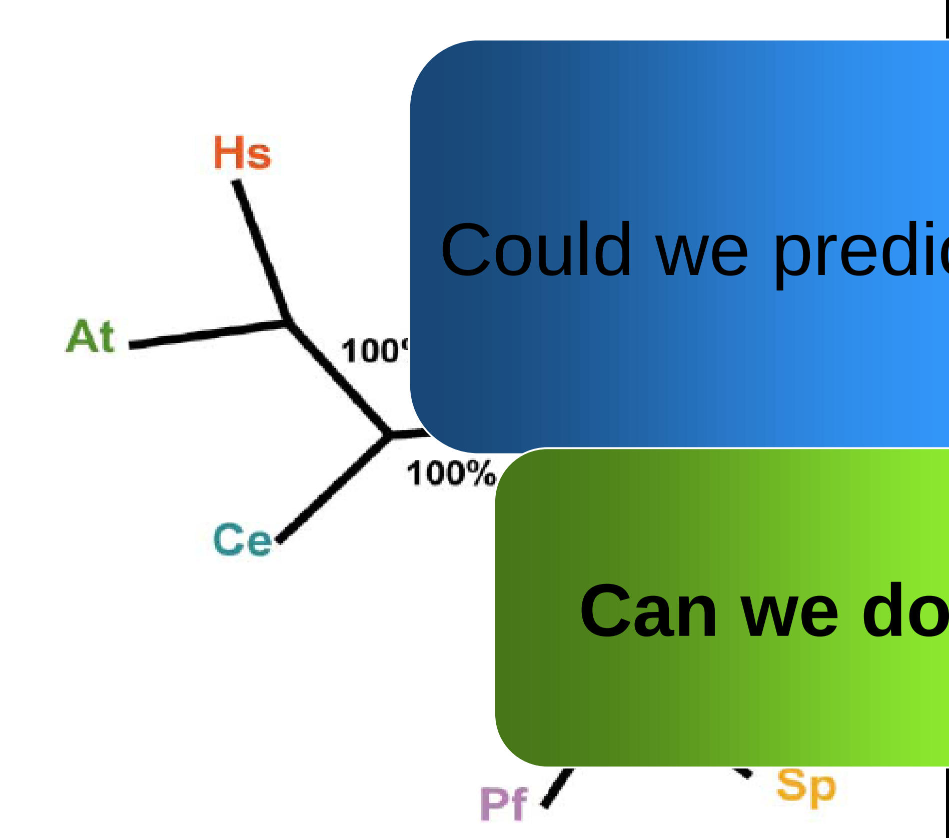


|  | 105 | 255 | 256 | 291 | 312 | 394 |
|---|---|---|---|---|---|---|
| Sc | 0 | 0 | 0 | 1 | 0 | 0 |
| Dr | 1 | 0 | 0 | 0 | 0 | 0 |
| Ar | 0 | 0 | 1 | 0 | 0 | 1 |
| Ce | 1 | 0 | 1 | 0 | 0 | 0 |
| Hs | 1 | 0 | 1 | 0 | 1 | 0 |
| Sp | 0 | 1 | 0 | 0 | 0 | 0 |
| Ag | 1 | 0 | 0 | 0 | 0 | 0 |
| Pf | 0 | 0 | 1 | 0 | 0 | 0 |

Pos. in the alignment

Intron starts at this position

# But… Dollo parsimony fails



Could we predict this will not work ?

**Can we do something about it?**

Figure 2. A Maximum Parsimony Tree Based on the Concatenated Intron Absence/Presence Data

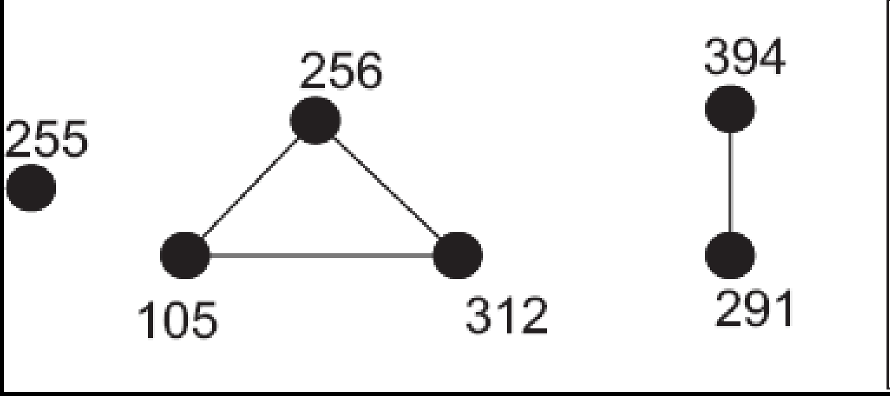
Only the data for conserved alignment regions were analyzed. The unrooted tree was constructed by using Dollo parsimony. Only one most parsimonious tree was obtained; the numbers at the interior branches are bootstrap values with 1000 replicates. The species abbreviations are as in Figure 1.

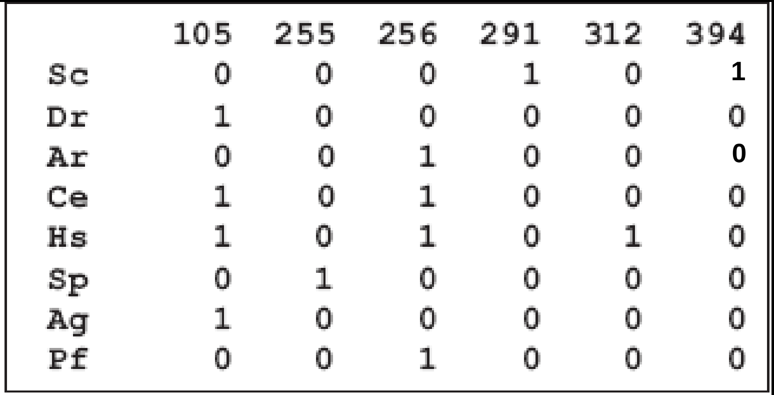
**Rogozin, Wolf, Sorokin, Mirkin, Koonin, 2003**

- Parsimony doesn’t work
- How about compatibility criterion?
- This doesn’t work for introns (we remove to much)
- Is there a weaker consistency measure that can be applied instead of compatibility?

# Character overlap graph

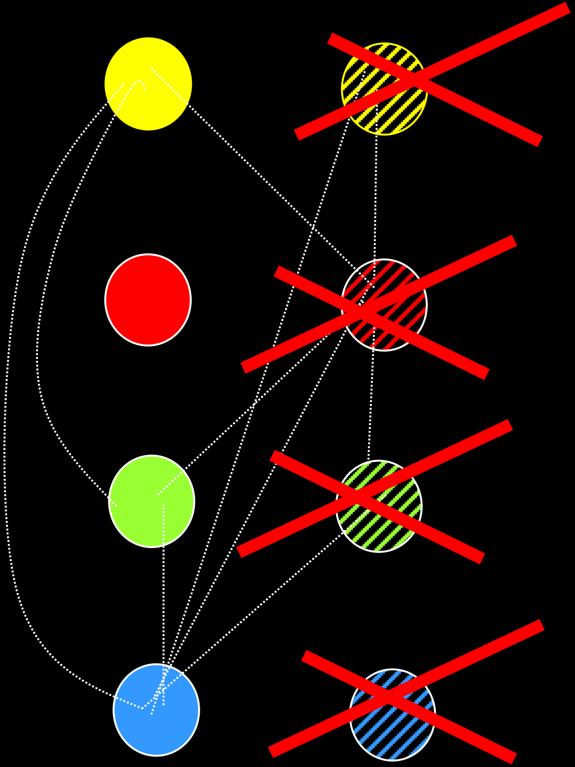
- Characters = nodes
- Two nodes are connected by an edge if there is a taxon which contains both characters (both characters have sate 1)





| | 105 | 255 | 256 | 291 | 312 | 394 |
|---|---|---|---|---|---|---|
| Sc | 0 | 0 | 0 | 1 | 0 | 1 |
| Dr | 1 | 0 | 0 | 0 | 0 | 0 |
| Ar | 0 | 0 | 1 | 0 | 0 | 0 |
| Ce | 1 | 0 | 1 | 0 | 0 | 0 |
| Hs | 1 | 0 | 1 | 0 | 1 | 0 |
| Sp | 0 | 1 | 0 | 0 | 0 | 0 |
| Ag | 1 | 0 | 0 | 0 | 0 | 0 |
| Pf | 0 | 0 | 1 | 0 | 0 | 0 |

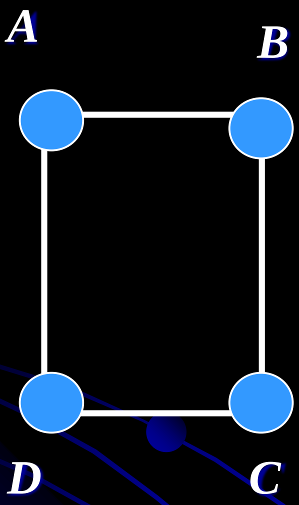
# Difference between character overlap graph and attribute overlap graph



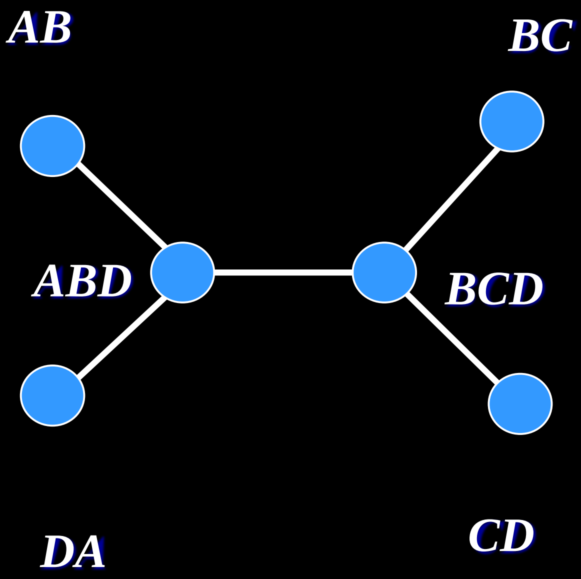
# New Concept: Persistent characters

**Assume set of taxa {AB, BC , CD, DA} where A,B,C,D characters**

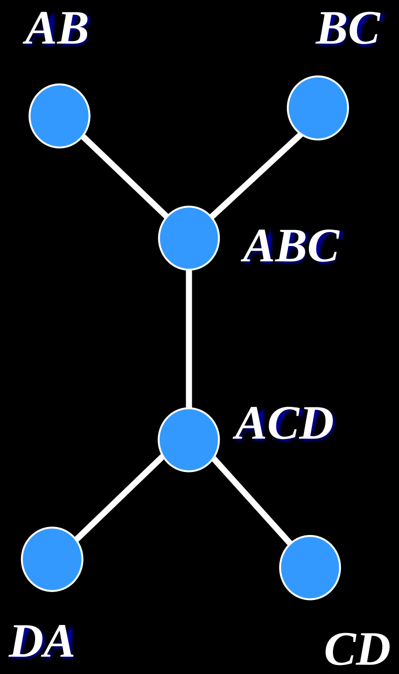
**Two possible tree topologies**



**Character overlap graph**



**B,D have to change sate twice**



**A,C have to change sate twice**

# Persistent characters

- A character is persistent if it does not belong to a hole.
- A set of characters is persistent if and only if the character overlap graph is chordal
- Property: a set of characters where each character can change its state at most twice (insertion first and then deletion) is persistent
- Thus persistency is a weaker assumption than compatibility

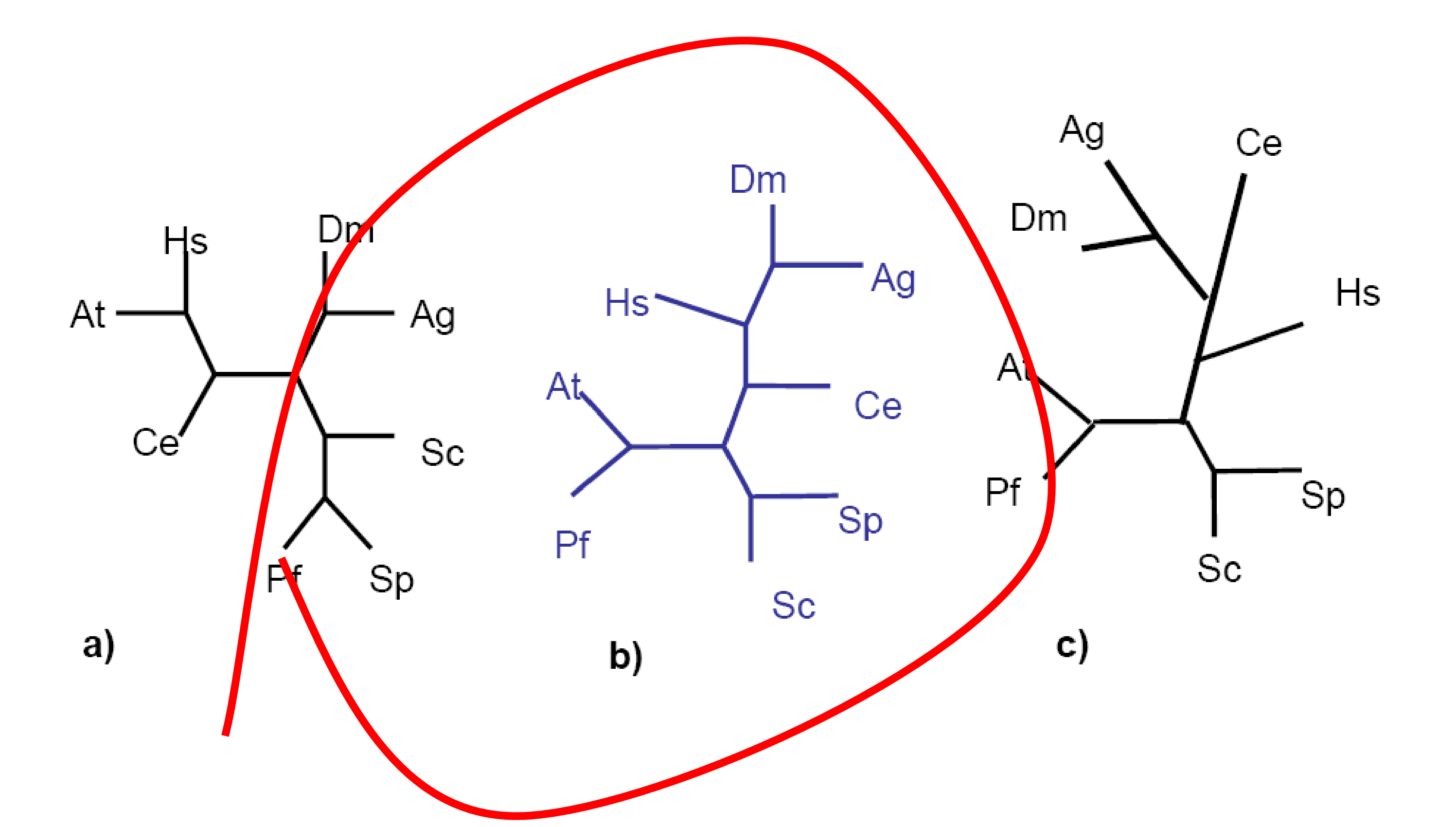
Przytycka *RECOMB 2006*

# Removing non-persistent characters

- Remove smallest number of character so that character overlap graph is chordal
- Construct the tree from the remaining data.
- Problem: Finding such minimal set is NP-complete; so is finding all holes.
- Heuristic approach: consider only squares and remove them in a greedy way.
- For the intron data, enough characters were preserved to build the tree

**Przytycka, *RECOMB 2006***

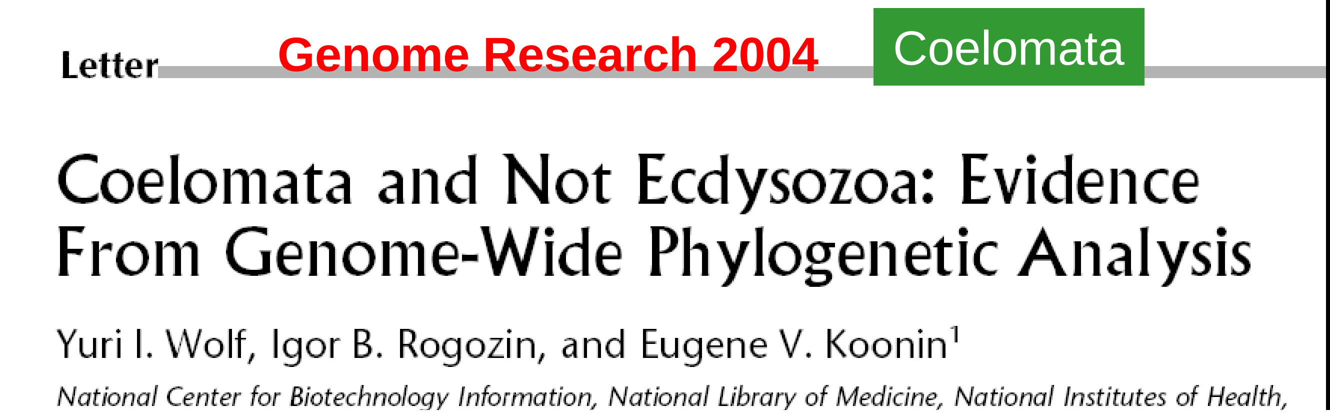
# Resulting Tree



*Coelomata* *Ecdysozoa*

**Przytycka, *RECOMB 2006***

# Coe



Letter

Genome Research 2004

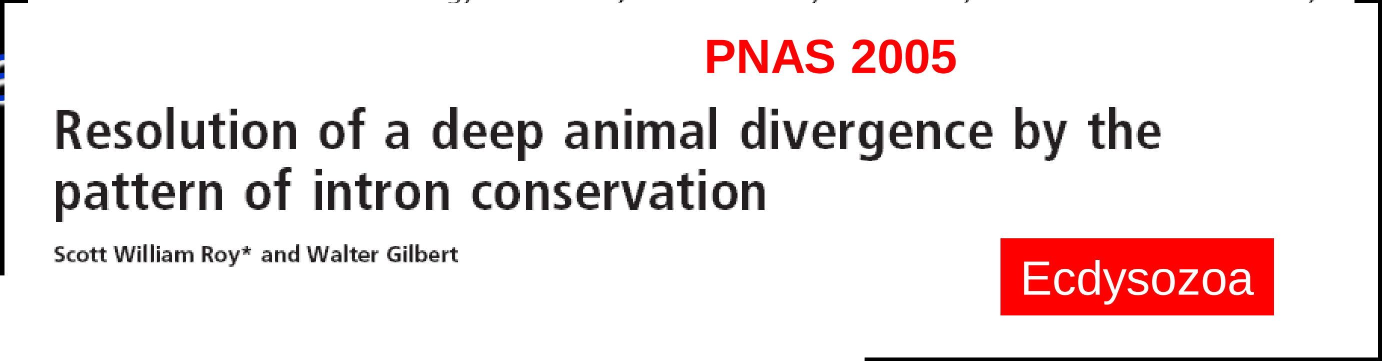
Coelomata

## Coelomata and Not Ecdysozoa: Evidence From Genome-Wide Phylogenetic Analysis

Yuri I. Wolf, Igor B. Rogozin, and Eugene V. Koonin[1]

*National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health,*

Coelomata



PNAS 2005

## Resolution of a deep animal divergence by the pattern of intron conservation

Scott William Roy* and Walter Gilbert

Ecdysozoa

Ecdys

Aguin

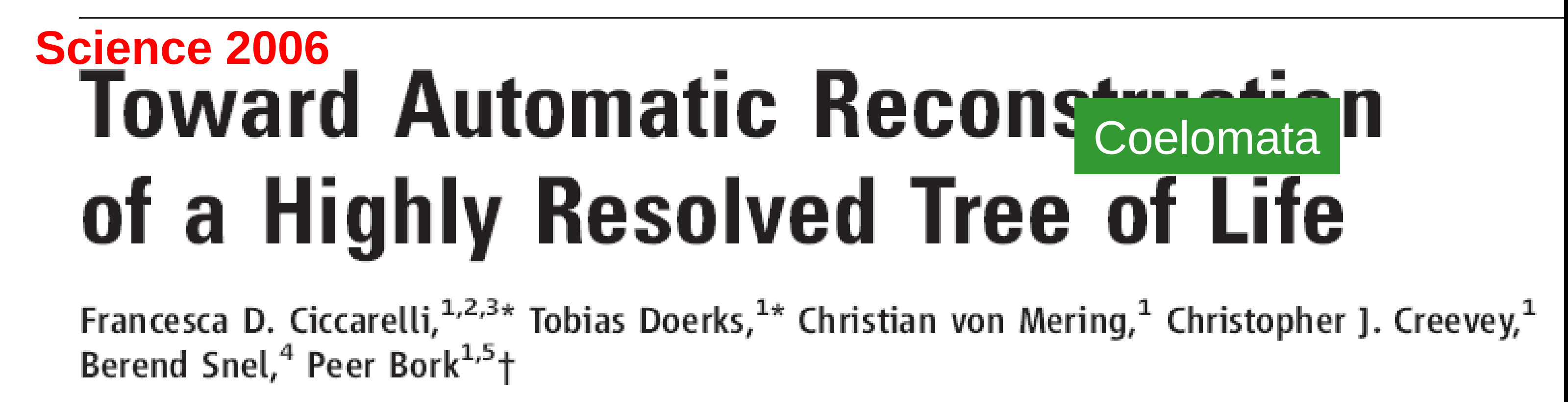
**Przytycka *RECOMB 2006***

Coelomata

Girbet et al. 2006

Pete

Malla



Science 2006

## Toward Automatic Reconstruction of a Highly Resolved Tree of Life

Coelomata

Francesca D. Ciccarelli,[1,2,3]* Tobias Doerks,[1]* Christian von Mering,[1] Christopher J. Creevey,[1] Berend Snel,[4] Peer Bork[1,5]†

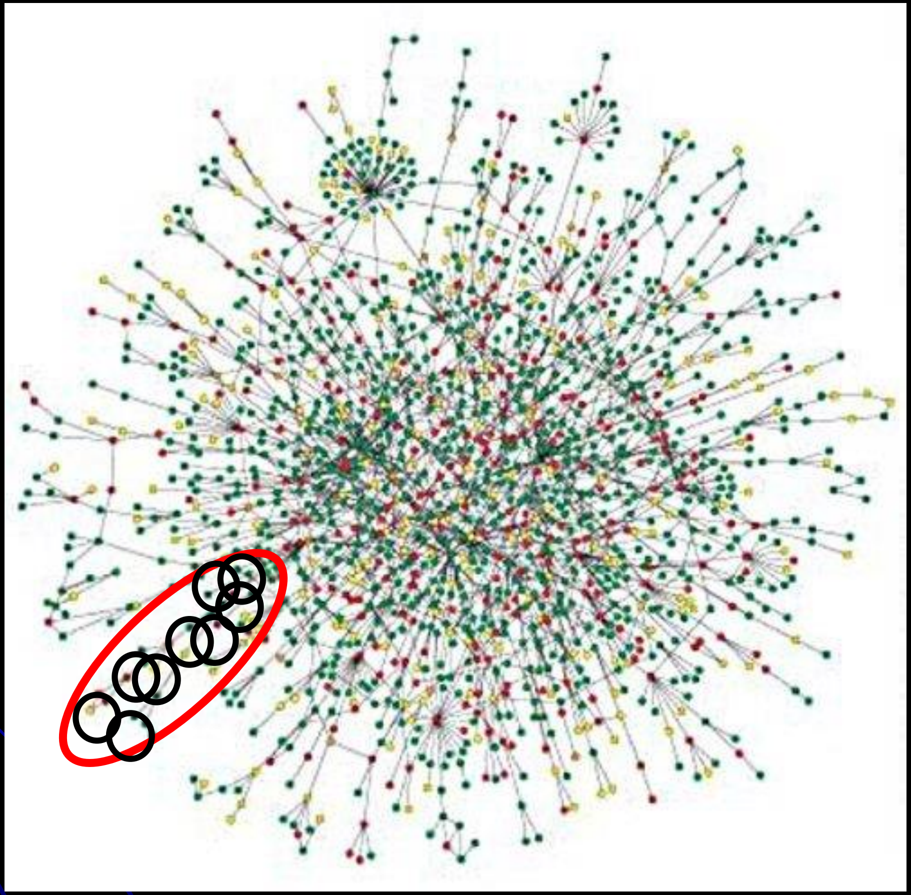
**Przytycka *RECOMB 2006***

# Is the number of holes correlated with the applicability of Dollo parsimony?

| Type of character overlap graph | Dollo applicable? | Number of squares in real data | Number of squares in the null model |
|---|---|---|---|
| domains | YES | 251 | 55,983 |
| introns | NO | 954 667 368 | 1389 751 510 |

**Przytycka *RECOMB 2006***

# Investigating protein-protein interaction networks

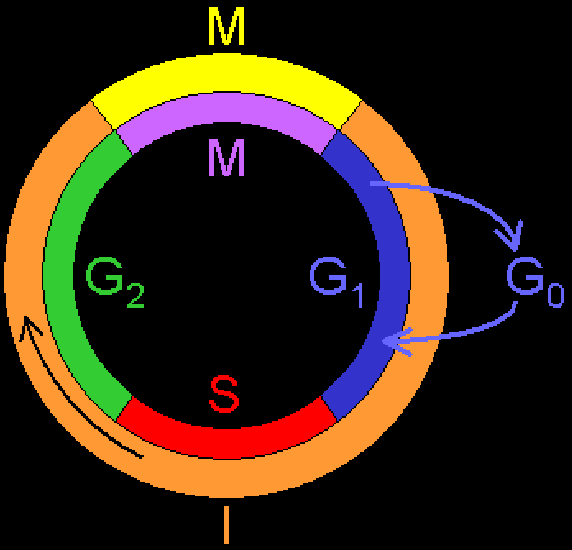


**Zotenko, Guimaraes, Jothi, Przytycka; *RECOMB 2005* (Sys. Biol)**
*Algorithms for Molecular Biology* 2006

# Functional Modules and Functional Groups

- Functional Module: Group of genes or their products in a metabolic or signaling pathway, which are related by one or more genetic or cellular interactions and whose members have more relations among themselves than with members of other modules (Tornow *et al.* 2003)
- Functional Group: protein complex (alternatively a group of pairwise interacting proteins) or a set of alternative variants of such a complex.
- Functional group is part of functional module

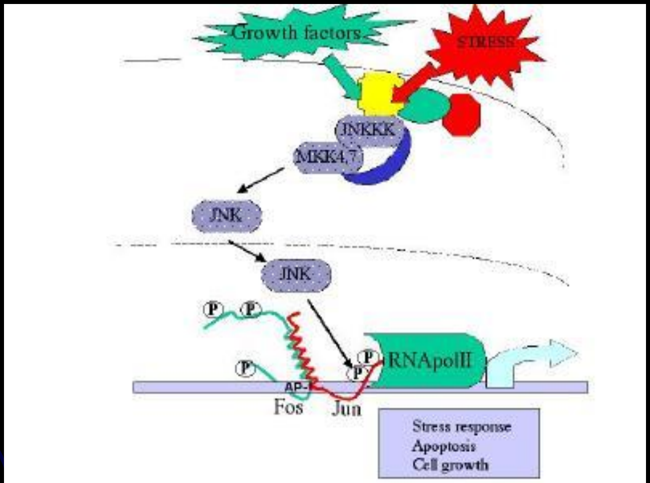
# Protein interactions are not static



Two levels of interaction dynamics:

- Interactions depending on phase in the cell cycle

- Signaling



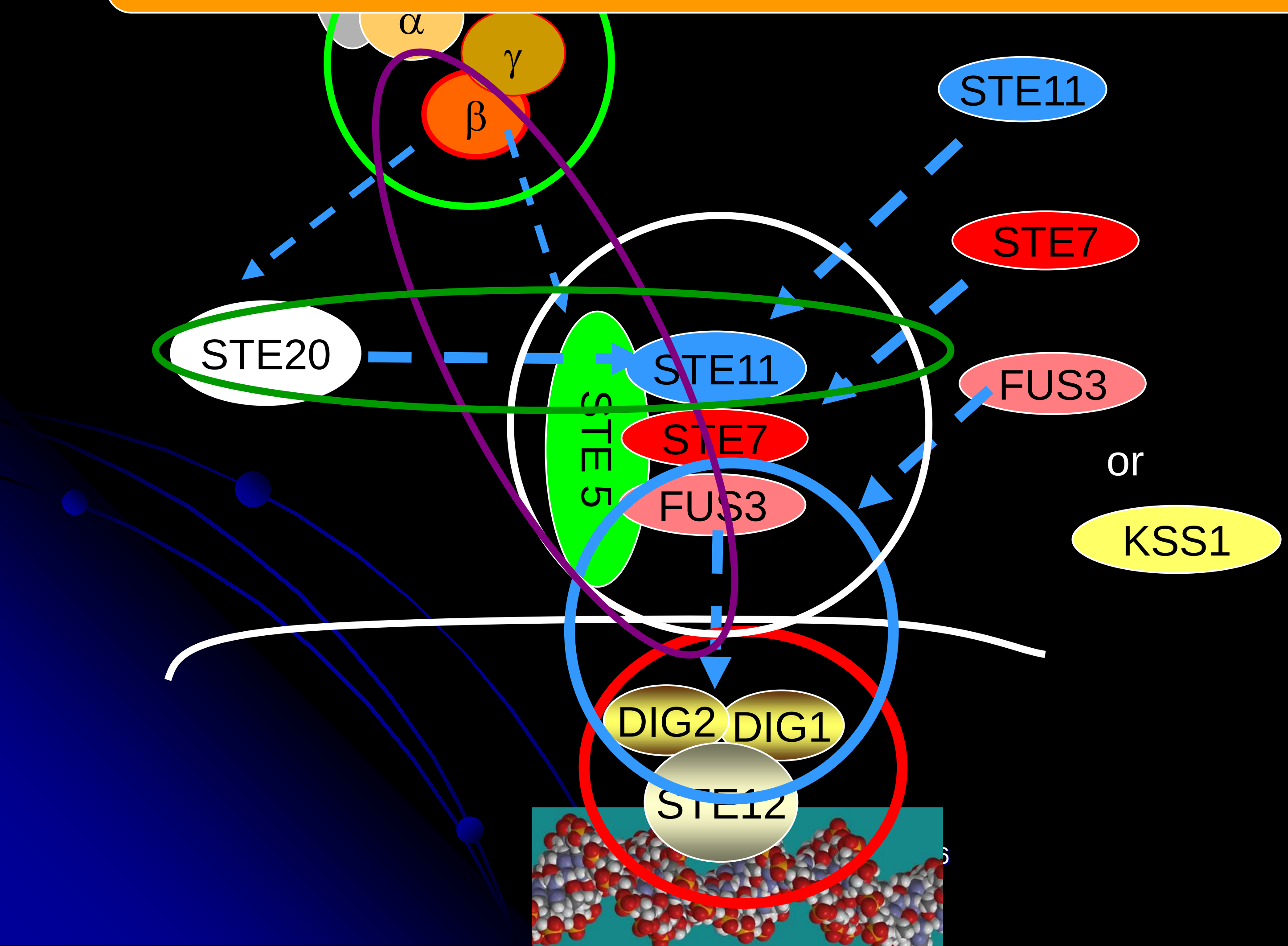
# Challenge

Within a subnetwork (functional module) assumed to contain molecules involved in a dynamic process (like signaling pathway), identify functional groups and partial order of their formation

Activation of the pathway is initiated by the binding of extracellular pheromone to the receptor

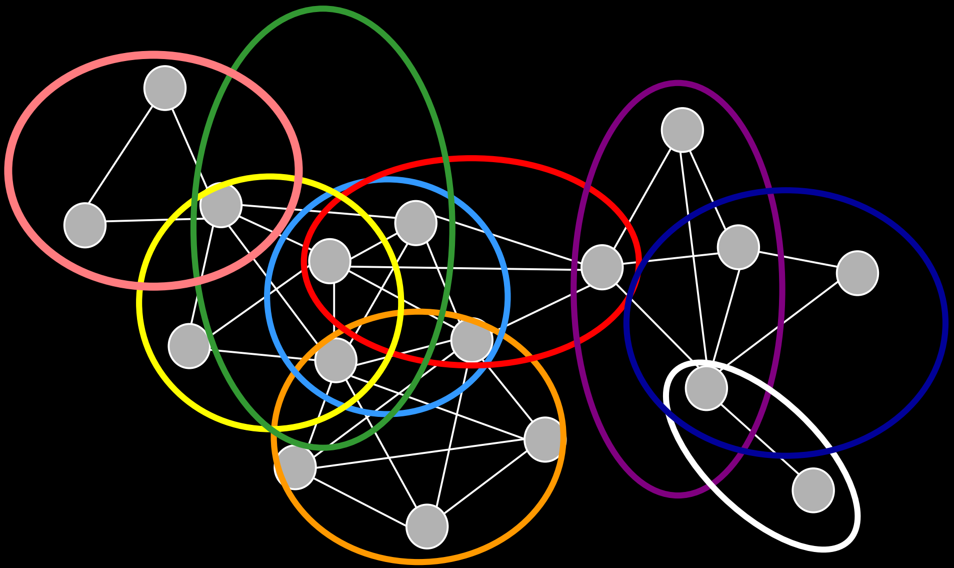
which in turn catalyzes the exchange of GDP for GTP on its cognate G protein alpha subunit Gα

G β is freed to activate the downstream MAPK cascade

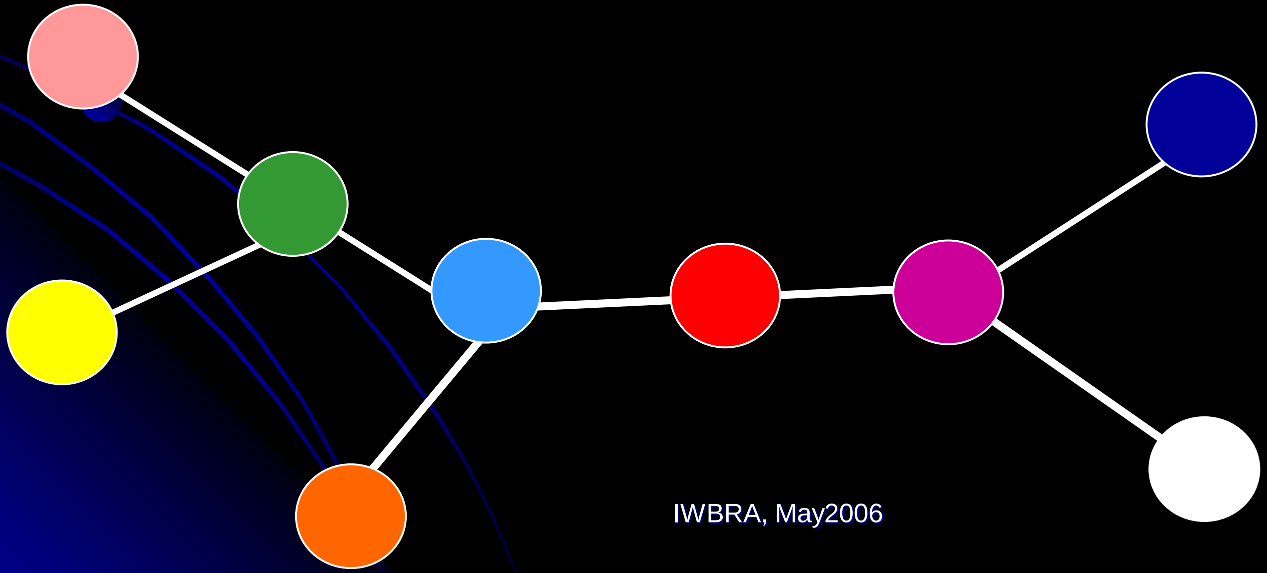


## Assume that a process satisfies the following properties:

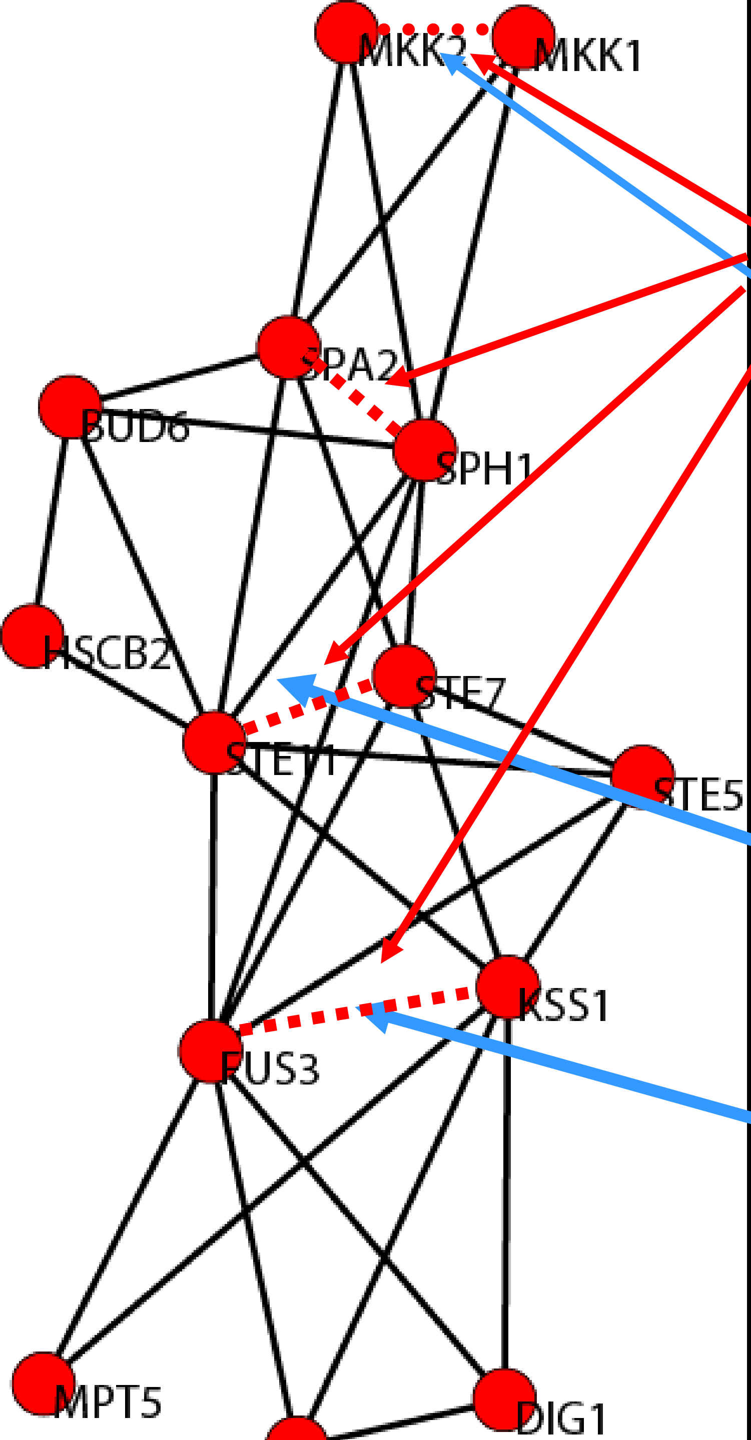
- Functional modules are maximal cliques
- Functional modules are formed according to some partial order
- Each protein enters the process once, participates is some consecutive steps and then leaves



# Clique tree



- Is protein interaction network chordal?
- Not really
- Consider smaller  subnetworks like functional modules
- Is such subnetwork chordal?
- Not necessarily but  if it is not it is typically close to it!
- Furthermore, the places where  they violates chordality tend to be of interest.



**Add special “OR” edges**

t

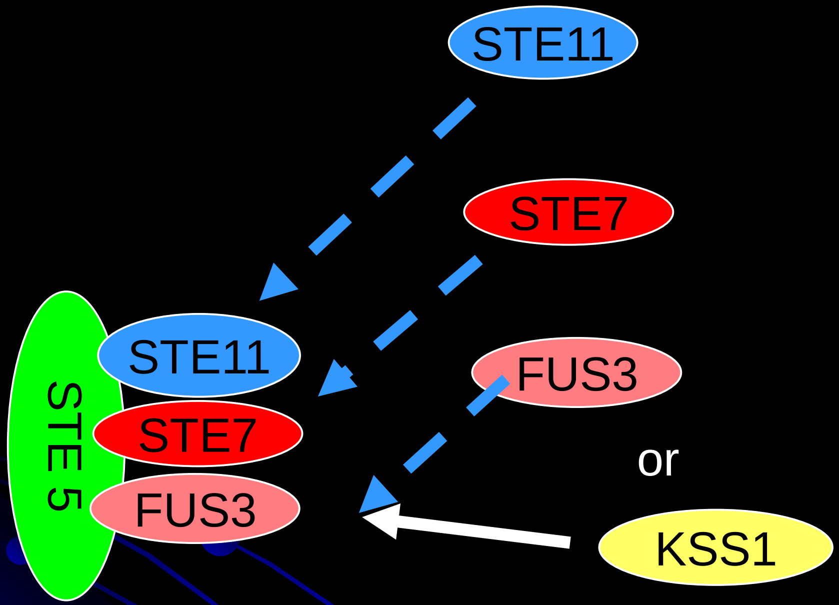
**assembled by Spirin *et al*. 2004**

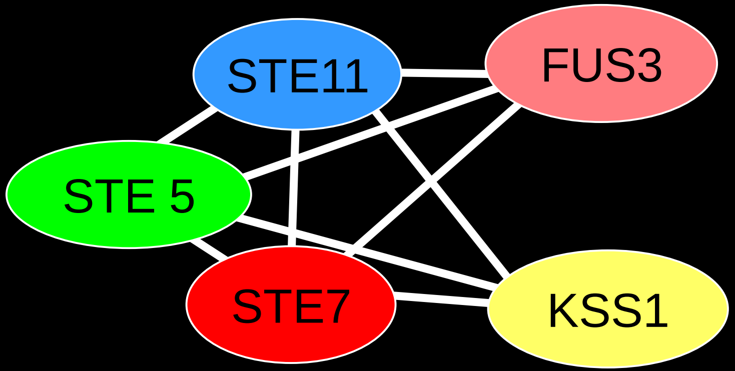
**Square 1:**
**MKK1, MKK2 are experimentally confirmed to be redundant**

**Square 2:**
**STE11 and STE7 –**
**missing interaction**

**Square 3:**
**FUS3 and KSS1 –**
**similar roles (replaceable but not redundant)**

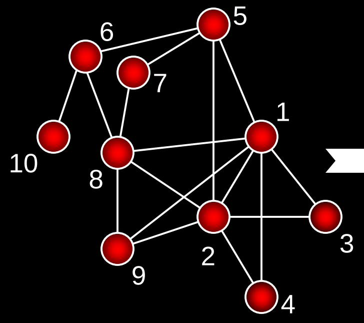
# Example: representing two variants of a complex





**STE5 ∧STE11∧ STE7 ∧(FUS3 v KSS1)**

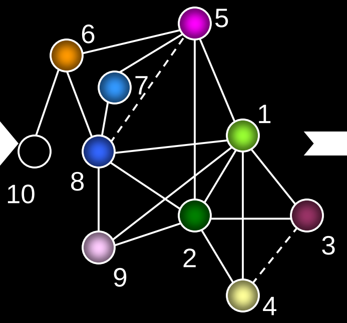
Original Graph, G



Graph modification

1. Add edges between nodes with identical set of neighbors
2. Eliminate *squares* (4-cycles) (if any) by adding a (restricted) set of “fill in” edges connecting nodes with similar set of neighbors

Modified Graph, G*



Is the modified graph chordal?

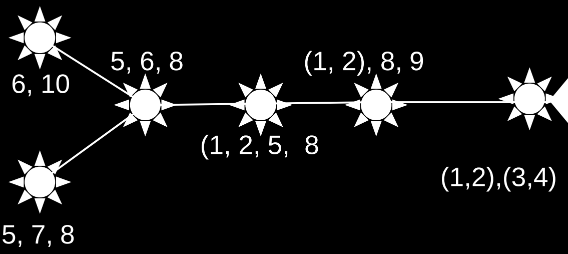
No

STOP

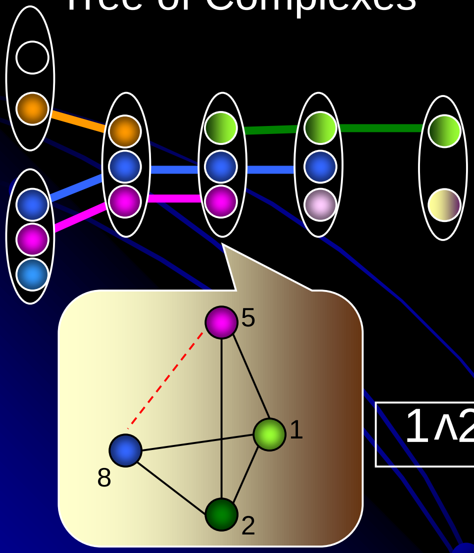
Yes

1. Compute *perfect elimination order* (PEO)
2. Use PEO to find maximal cliques and compute *clique tree*

Maximal Clique Tree of G*



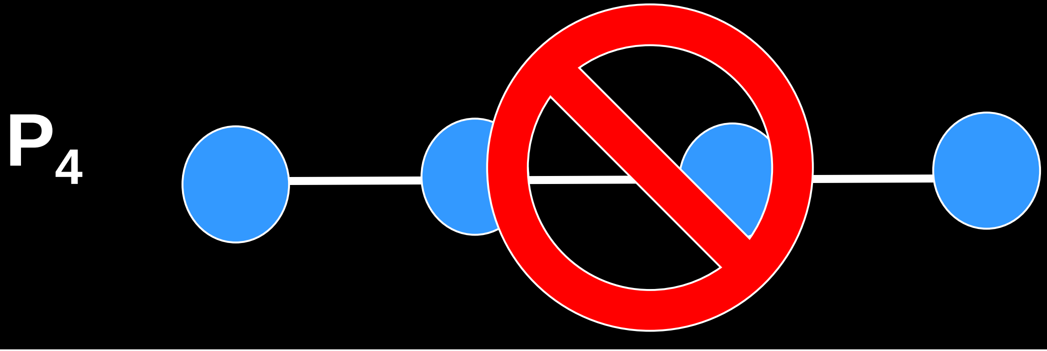
Tree of Complexes



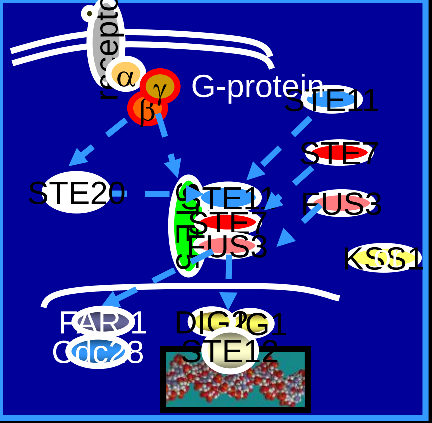
1∧2∧(5v8)

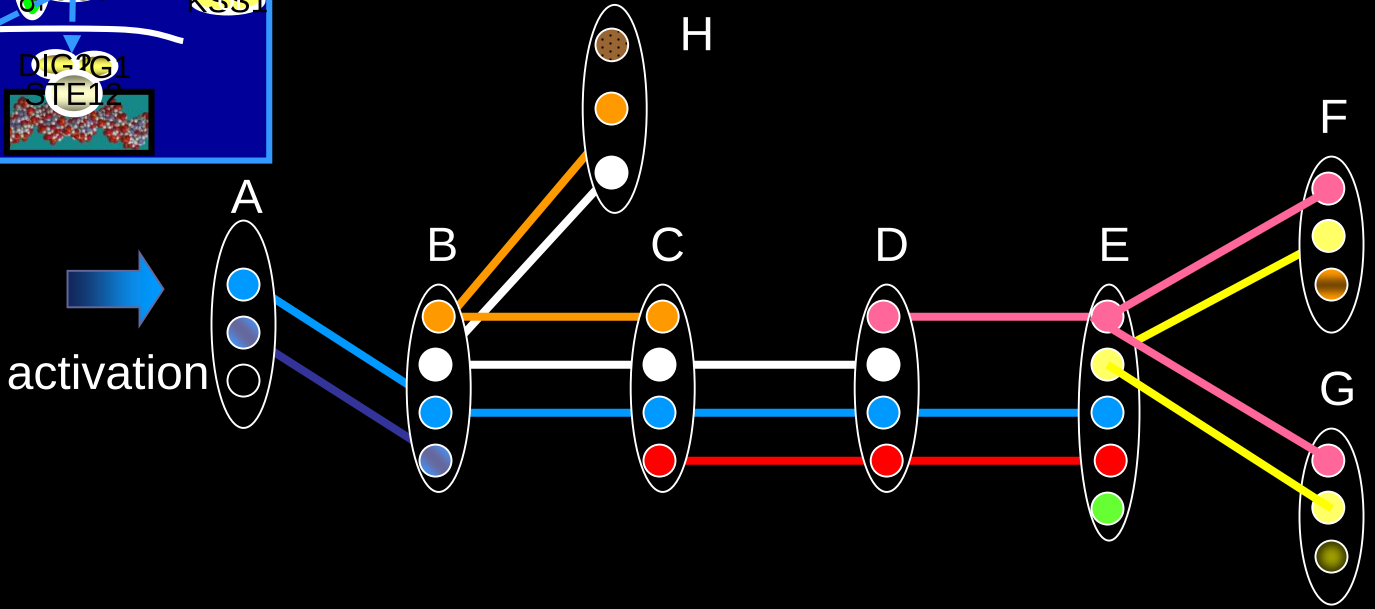
● Protein – – – Fill-in edge ✹ Maximal clique

# Not all graphs can be represented by Boolean expression



Cographs = graphs which can be represented by Boolean expressions





| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 🟤 | = MKK1 v MKK2 | 🔵 | = STE11 | ● | = FUS3 | ○ | = HSCB2 |
| ⚪ | = SPH1 | 🟢 | = STE5 | 🟡 | = KSS1 | ● | = BUD6 |
| 🟠 | = SPA2 | 🔴 | = STE7 | 🟤 | = DIG1Λ DIG2 | ● | = MPT5 |

***FUNCTIONAL GROUPS***

***A = HSCB2 Λ BUD6 Λ STE11***

***C = (SPH1 v SPA2) Λ(STE11 v STE7)***

***E = STE5 Λ(STE11 v STE7) Λ(FUS3 v KSS1)***

***G = (FUS3 v KSS1) ΛMPT5***

***B = BUD6 Λ(SPH1 v SPA2) ΛSTE11***

***D = SPH1Λ(STE11 v STE7) Λ FUS3***

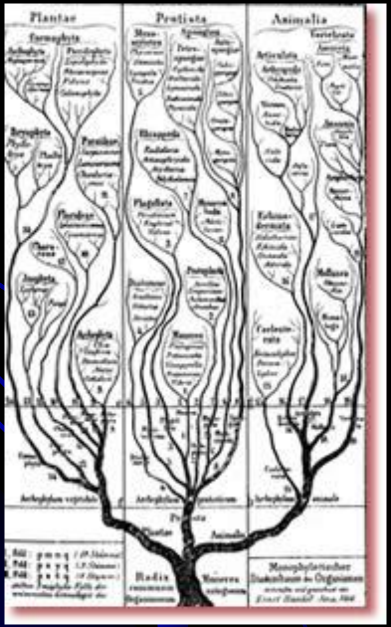
***F = (FUS3 v KSS1) ΛDIG1Λ DIG2***

***H = (MKK1 v MKK2)Λ(SPH1 v SPA2)***

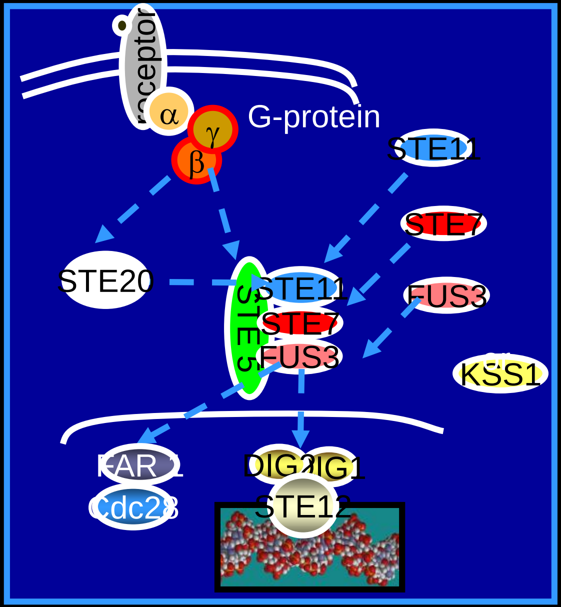
# Summary

- Chordal graphs can be used naturally in modeling biological processes
  - Persistency analysis
  - Delineating protein complexes and their overlap analysis

evolutionary



molecular

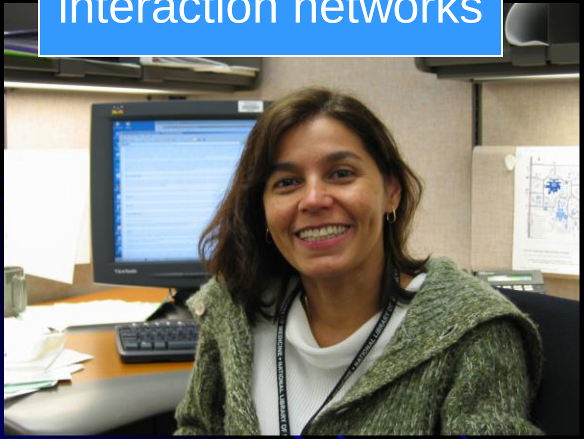


Protein domains:
In collaboration with
Dannie Durand, CMU

# Thanks

- Funding: NIH intramural program, NLM
- Przytycka's lab members:

Analysis of protein interaction networks



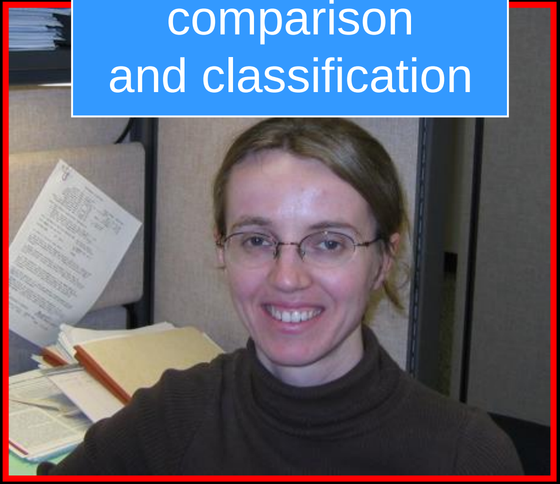
Katia S. Guimarães

(visitor)

Orthology clustering, Co-evolution

*Raja Jothi*

Protein Complexes
Protein structure: comparison and classification



*Elena Zotenko*

# Protein domains

DOMAINS:

- Building blocks for large proteins.
- Evolutionary units.
- Can fold independently and carry some specific function

# Domain level evolution

## Assumptions

- Protein architecture is described by the set of its domains (we ignore the order)
- Operations: **insertion** and **deletions**

Domains typically correspond to functional

Inferring an ancestral architecture that contains two domains never observed together

Given a family of multidomain proteins, character overlap graph is chordal if and only if each domain pair that is inferred to belong to same ancestral architecture

Persistency is a reasonable assumption for protein domain evolution

# Is character overlap graph for multidomain proteins chordal?

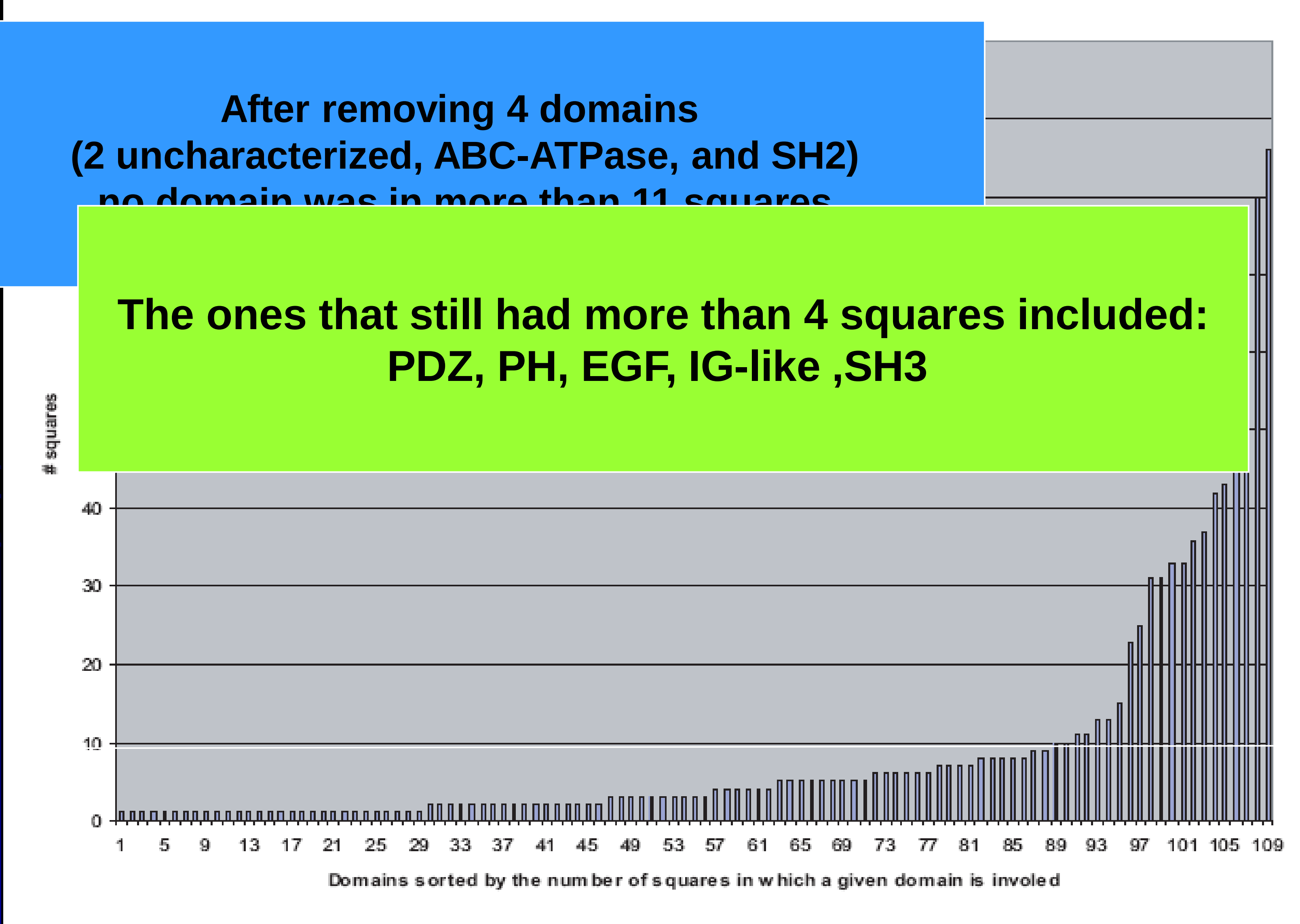
| *n** | # families | %PP | %SDP | %CDP | Random graphs | |
|---|---|---|---|---|---|---|
| | | | | | Uniform | Degree preserving |
| 4-5 | 143 | 57 | 99 | 99.5 | 80 | 98 |
| 6-8 | 130 | 37 | 99 | 100 | 31 | 66 |
| 9-10 | 40 | 28 | 100 | 100 | 17 | 25 |
| 11-20 | 104 | 13 | 87 | 99 | 1.7 | 1.0 |
| 21-30 | 34 | 6 | 53 | 88 | 0 | 0 |
| ≥30 | 28 | 0 | 15 | 50 | 0 | 0 |

*34 superfamilies do not safisfy CDP, including TyrKc, Ig, PH, EGF, CUB, SH3, C1, Myosin_Tail*

**n* is the number of distinct domains in the superfamily.*

**Przytycka, Davis, Song, Durand *RECOMB 2005***

# Domains involved in large number of squares: promiscuity profile

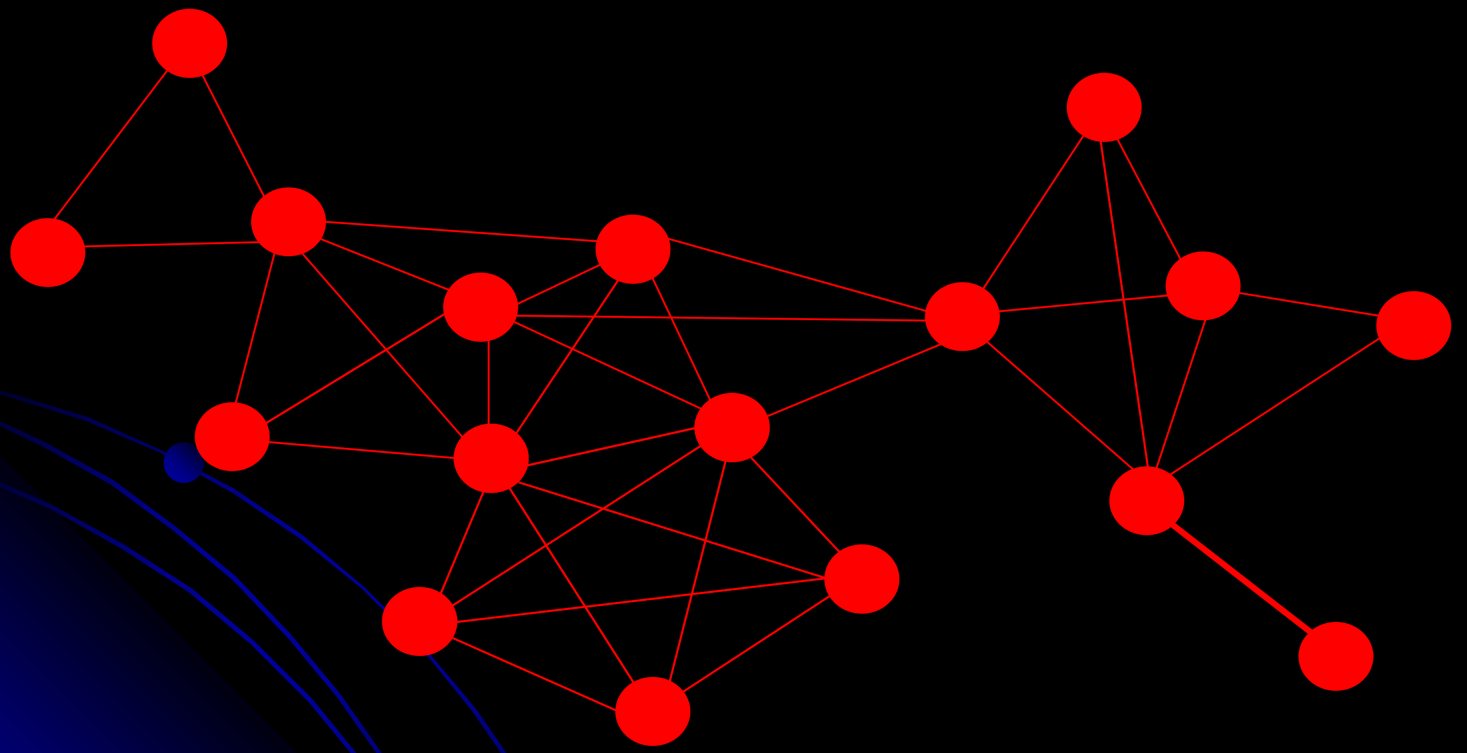


After removing 4 domains
(2 uncharacterized, ABC-ATPase, and SH2)
no domain was in more than 11 squares

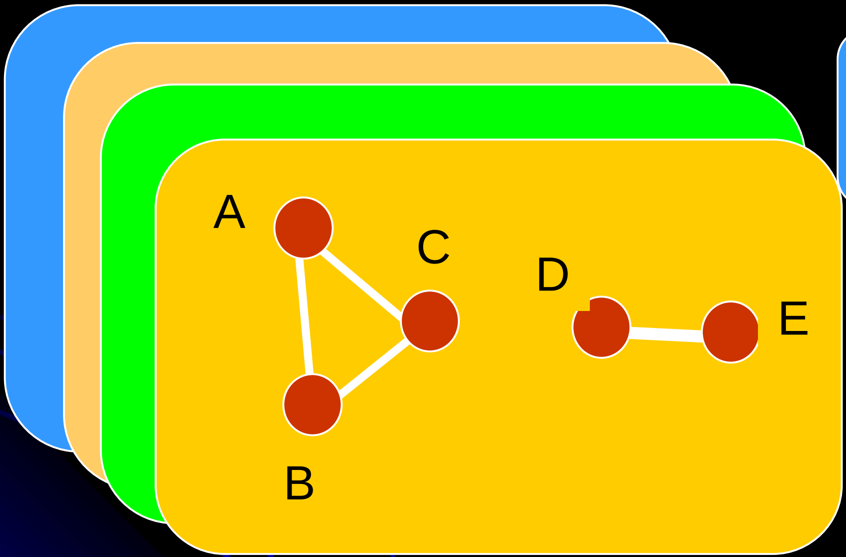
The ones that still had more than 4 squares included:
PDZ, PH, EGF, IG-like ,SH3

# Overlaps between Functional Groups

**For an illustration functional groups = NOT maximal cliques**



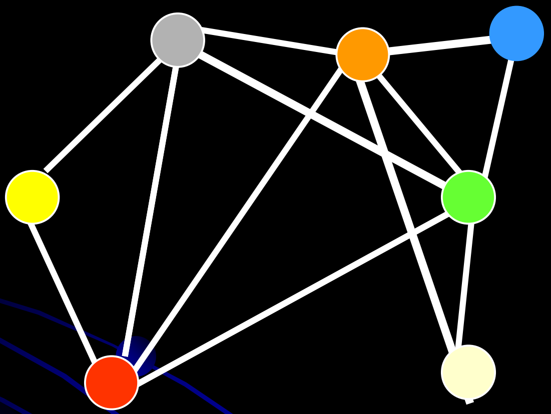
# Representing a functional group by a Boolean expression

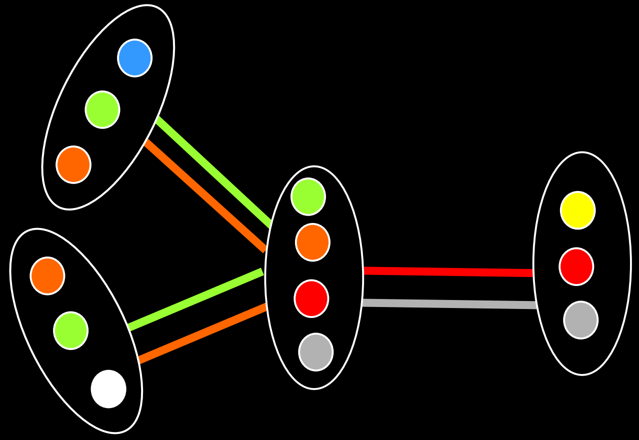


$(A \wedge B \wedge C) \vee (D \wedge E)$

Assume that a process satisfies the following properties:

- **Functional modules are formed according to some partial order**
- **each protein enters the process once, participates is some consecutive steps and then leaves**





**Clique tree representation** :

Nodes = functional groups

Edges = possible partial order of their formation